SEPRES: Sepsis prediction via a clinical data integration system and real-world studies in the intensive care unit

Qiyu Chen[#], Ranran Li[#], ChihChe Lin, Chiming Lai, Dechang Chen, Hongping Qu, Yaling Huang, Wenlian Lu[*], Yaoqing Tang[*], Lei Li[*]

Affiliations

Department of Applied Mathematics, Fudan University, Shanghai, China (Q Chen BSc, Prof W Lu PhD)

Department of Critical Care Medicine, Ruijin Hospital, Shanghai Jiaotong University School of Medicine, Shanghai, China (R Li PhD, D Chen MD, PhD, H Qu MD, PhD, Y Tang MD, PhD, L Li MD, PhD)

Shanghai Electric Group Co., Ltd. Central Academe, Shanghai, China (C Lin PhD, C Lai MA, Y Huang BEc)

Correspondence to:

Dr Lei Li, Department of Critical Care Medicine, Ruijin Hospital, Shanghai Jiaotong University School of Medicine, Shanghai, 200025, China, lileiys1023@yeah.net

Dr Yaoqing Tang, Department of Critical Care Medicine, Ruijin Hospital, Shanghai Jiaotong University School of Medicine, Shanghai, 200025, China, tangyaoqing@126.com

Dr Wenlian Lu, Department of Applied Mathematics, Fudan University, Shanghai, 200433, China, wenlian@fudan.edu.cn

Summary

Background: Sepsis is vital in critical care medicine, and early detection and intervention are key to survival. We aimed to establish an early warning system for sepsis based on a data integration system that can be implemented in the intensive care unit (ICU).

Methods: We trained the LightGBM and multilayer perceptron on the open-source database Medical Information Mart for Intensive Care for sepsis prediction. An ensemble sepsis prediction model was established based on the transfer learning and ensemble learning technique on the private dataset of Ruijin Hospital. The Shapley Additive Explanations analysis was applied to present feature importance on the prediction inference. With the development of data-integrating hub to collect and transmit data from different brands of ICU medical devices, the data integration system was established to receive, integrate, standardize, and store the real-time clinical data. In this way, the sepsis prediction model developed in the ICU of the Ruijin

[#] Contributed equally as first authors

* Contributed equally as senior authors

Hospital for the real-world study of sepsis early warning on ICU management. The trial was registered with ClinicalTrials.gov (NCT05088850).

Findings: Our best early warning model achieved an area under the receiver operating characteristic curve (AUC) of 0·9833 in the task of detecting sepsis in 4-h preceding on the open-source database, while our ensemble model achieved an AUC of 0·9065–0·9436 in the retrospective research from 1–5-h preceding on the private database, and 0·8636–0·8992 in real-time real-world studies using the data integration system in the ICU of the Ruijin Hospital. In the continuous early warning process of patients admitted to the ICU, 22 patients who met the diagnostic criteria for sepsis during hospitalization were predicted as positive cases; 29 patients without sepsis were predicted as negative cases. Additionally, 17 patients were predicted as false-positive cases; in six patients with sepsis during ICU stay, the predicted probabilities at different time nodes were all less than the warning threshold 0·7 and predicted as false-negative cases.

Interpretation: Machine learning models could allow accurate and real-time inference to detect sepsis onset within 5-h preceding at most with the help of the data integration system. We identified the features such as age, antibiotics, ventilation, and net balance to be important for the sepsis prediction inference. We argue that this system has promising potential to improve ICU management by helping medical practitioners identify at-sepsis-risk patients and prepare for timely diagnosis and intervention.

Funding: Shanghai Municipal Science and Technology Major Project, the ZHANGJIANG LAB, and the Science and Technology Commission of Shanghai Municipality.

Introduction

Sepsis, an infection-induced syndrome of physiological, pathological, and biochemical abnormalities, is a global healthcare issue that is associated with unacceptably high mortality and long-term morbidity among patients in an intensive care unit (ICU),[1,2] and is responsible for a substantial cost burden on health care resources.[3] Early detection and timely administration of appropriate antibiotics may be the most important factors to improve the prognosis of patients with sepsis.[4] However, nonspecific symptoms of patients with sepsis lead to delayed diagnosis and delayed intervention.[5]

Machine learning has emerged as a promising tool for early detection of sepsis occurrence through intensive management based on electronic medical records, laboratory data, and biomedical signals.[6,7] In 2016, Singer et al. proposed a new definition (Sepsis-3) of sepsis.[2] According to this, many recent studies on sepsis prediction defined sepsis by Sequential Organ Failure Assessment (SOFA) and infection instead of Systemic Inflammatory Response Syndrome (SIRS).[8-12] Prospective studies have shown that implementation of machine learning-based sepsis prediction algorithms in hospitals can reduce in-hospital mortality and length of stay.[13,14] In addition, many machine learning models provide superior model performance at the cost of transparency and interpretability, which has become a barrier to clinical application. Algorithms based on gradients, attention mechanism, and Shapley values are used to interpret the machine learning models.[15-17]

Most studies on sepsis detection used historical medical data, such as the Medical Information Mart for Intensive Care (MIMIC).[18] However, the implementation of the detection model in the ICU for real-time prediction is complex. The raw data needed for model inference, such as

bedside data, laboratory data, demographic data, and doctor's orders, usually come from different devices. Moreover, the information cannot interact directly due to differences in the data transfer protocols between devices. Efforts have been made to integrate bedside medical devices.[19-21] However, these data integration systems integrate a more limited number of devices and data types to present the complete perspective of a doctor. Moreover, they were mainly focused on data collection and presentation and lacked further functionality, such as real-time alerts. Meanwhile, previous studies on the prediction of sepsis were mainly retrospective, and prospective studies used only relatively simple variables, deployment methods, and models.

In this study, we aim to develop a data integration system for IntelliVue Information Center, Ventilators, Philips ICCA system, Laboratory Information System (LIS), and Hospital Information System (HIS), an ensemble machine learning model for the prediction of sepsis based on Sepsis-3 and establish a real-time early warning system for sepsis in the ICU, named SEpsis PREdiction System (SEPRES). In this way, we have developed the SEPRES in the ICU in the Ruijin Hospital and conducted real-world studies to analyze the performance of this system in the management of ICU patients.

Methods

SEPRES includes a data integration system equipped with a sepsis early warning module. The data integration system can collect, store, process, and display medical data. These functions were completed through the data integration machine, physical server, and network server. The sepsis early warning module included a sepsis prediction model and an interpretative tool. The sepsis prediction model is an ensemble of multiple machine learning models and utilizes the transfer learning technique to predict sepsis. The interpretative tool provides information on how the model works by assigning importance to the input features. Our research was approved by the Ruijin Hospital Ethics Committee (No. 2020 [140]).

Medical device integration hub

We developed a medical device integration hub that can acquire and transmit data from different brands of medical devices. The medical device integration hub consists of customized device connection lines, a hub, and an integrated data receiver. The identification module containing encoding is inserted into each medical device, enabling the hub to identify the type of online device and collect data automatically according to the communication protocol. The integrated data receiver receives and translates the raw data and uploads them to the integration server through the local area network. The medical device integration hub has the following functions:

- Device online services: Detect device connections and start a data reading program corresponding to the device.
- Decoding: Parsing raw data into structured data for further processing.
- Storage: Storing parsed data in native memory.
- Remote Settings: support remote system setup and send system status.
- Uploading: Upload data received to the specified database.

Details of the data extraction can be found in Appendix I.

The framework of a data-integrating system

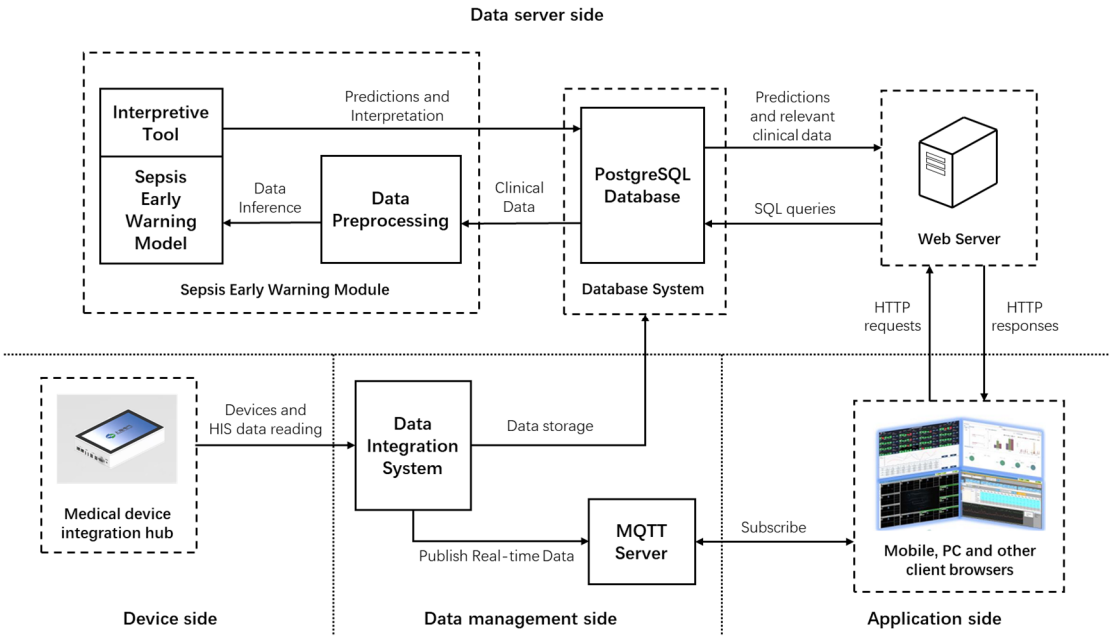


Figure 1 System Deployment Framework. The web release system of the sepsis prediction system (SEPRES) applies browser/server architecture.

As shown in Figure 1, the system includes a physical server with the PostgreSQL database to store the sepsis warning data, and the webserver deploys the system's user access portal. The architecture can be divided into the following parts.

- Device side: The medical device integration hub transmits the device and HIS data to the data integration system through the local area network.
- Data management side: Heterogeneous data are integrated into the data integration system. The interface data, service data, and model predictions are stored and managed by the Structured Query Language (SQL) server, while the parts needed for the sepsis early warning module are sent to the PostgreSQL database. The Message Queuing Telemetry Transport (MQTT) server sends real-time data from the data integration system to the browser.
- Data server side: The web server responds to the browser's request and calls the sepsis early warning module. Data preprocessing and model inference are then executed, and the predictions are stored in the PostgreSQL database. The data server side includes some related services (real-time calculation of the SOFA score, determination of suspected infection, data statistics, data charts, and historical data query).
- Application side: The user's request is passed to the webserver in this layer, and the processing results are displayed in the system. The Java Script program is used for dynamic HTML page development, and the AJAX interface is used for data interaction with the webserver. Spring MVC is used to build full-featured MVC modules for web applications, combined with NODEJS to provide an elegant and highly maintainable method for creating templates. Users can use the system anytime and anywhere with a browser in various ways, such as PCs and mobile terminals.

System deployment

Figure 2 shows the medical device integration hub installed at Ruijin Hospital. The hub was placed at the bedside, receiving data from multiple devices via different interfaces shown at the

bottom of the figure, storing the last 72 h of data in native memory, and transmitting data with a time delay of less than 10 s. Interfaces distributed on the two sides of the hub include two universal network interfaces, four USB interfaces for the mouse, keyboard, and U disk, two HDMI and one VGA for extended display, and eight or 16 USB and Ethernet multiplexing interfaces for medical devices. The hub can integrate data from the monitor, ventilator, infusion pump, and dialysis machine. The processed data are then transmitted to the data integration system.

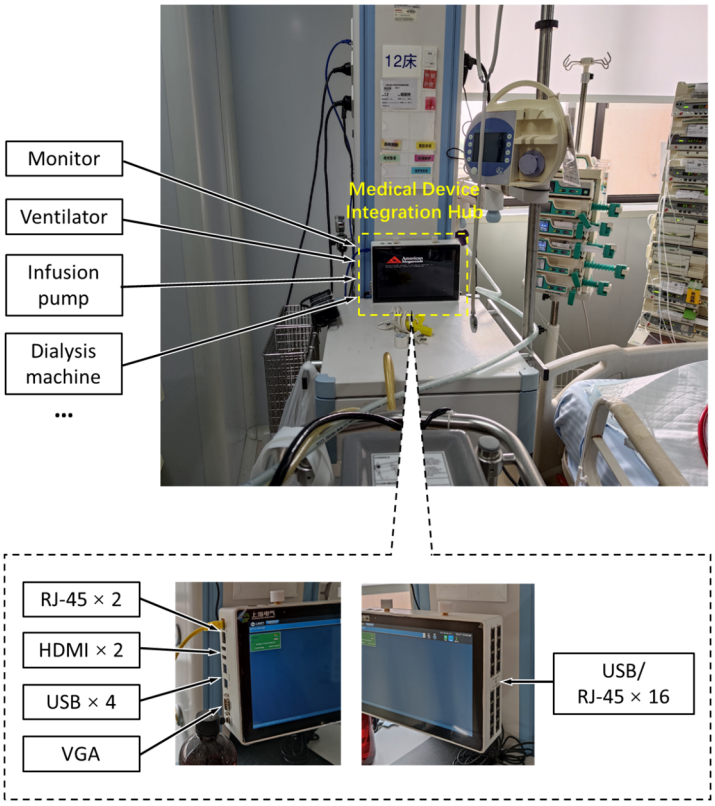


Figure 2 Medical device integration hub

Sepsis prediction model

Our goal was to develop a sepsis prediction model that can run in real-time in hospitals. To avoid insufficient data in the specific hospital for training, we first trained the models in the open-source database MIMIC and then retrained them in private hospital databases using transfer learning techniques to improve the performance. The final sepsis prediction model was obtained by integrating multiple transferred models using ensemble learning techniques.

Data acquisition

Data sources and inclusion criteria. Our study used the MIMIC-III database (version 1·4)[18] and the private Ruijin Hospital historical (RJ) database. MIMIC encompasses approximately 40,000 patients admitted to the ICU at Beth Israel Deaconess Medical Center in Boston from 2001 to 2012. Two tasks were established: inference on the MIMIC dataset by models trained on the MIMIC dataset and inference on the RJ dataset by models trained on the MIMIC and RJ datasets. The first task was to facilitate comparison with other articles, and the second was to apply the models clinically in Ruijin Hospital.

Patients who met all the following criteria were included in the case group:

1) At least 14 years old.

2) Sepsis onset at least 5 h after admission to the ICU.
3) Sepsis onset is the first instance since admission to the hospital.
Patients who met all the following criteria were included in the control group:
1) At least 14 years old.
2) Patients who stayed in the ICU for at least 5 h and have not had sepsis at this time.
3) Patients without ICD-9 codes for sepsis (785·52, 995·91, and 995·92).
4) SOFA score changes of no more than 1 point in an arbitrary continuous 72 h in the ICU stay.
The third criterion was excluded from the RJ database because ICD-9 codes were not recorded.
Sepsis-label definitions. Patients were followed throughout their stay in the ICU until discharge or development of sepsis according to the definition of the Third International Consensus for sepsis (Sepsis-3).[2] Specifically, if the timestamp of antibiotics (t_{abx}) and blood cultures ($t_{culture}$) meet the condition $t_{abx} - 24\ h \leqslant t_{culture} \leqslant t_{abx} + 72\ h$, the earlier timestamp of t_{abx} and $t_{culture}$ is defined as the timestamp of suspected infection (t_{sus}). The SOFA score was evaluated per hour within the time window $[t_{sus} - 48\ h, t_{sus} + 24\ h]$. The first hour with two or more points of increase in the SOFA score than the lowest prior score is defined as the onset of sepsis (t_{onset}).
Feature extraction. We extracted 78 and 63 patient variables from the MIMIC and RJ databases, respectively. After data cleaning, we extracted these variables as features, i.e., maximum, average, median, and minimum, at hourly intervals, and the missing data were padded by the nearest value before or a preset default value. After filtering, we obtained the MIMIC dataset with 1057 positive and 5834 negative episodes and the RJ dataset with 115 positive and 239 negative episodes. We used a 5-h time window from the episodes to predict sepsis. These two datasets were divided into training, validation, and test sets. See Appendix II for further details.

Prediction model

Machine learning model. Multiple models were tested on the MIMIC dataset, including support vector machine (SVM), multilayer perceptron (MLP), gradient boosting machine (GBM), and long short-term memory (LSTM). For GBM, we used XGBoost[22] and LightGBM[23] as implementations. A detailed introduction to these models can be found in Appendix V.
Training method. Some redundant features were removed to accelerate SVM and MLP training in the tasks. Data were standardized (i.e., each feature's value range was linearly scaled between 0 and 1) before training to eliminate magnitude differences between features. The hyperparameters and structures of each model were tuned according to the effects on the validation set. See Appendix VI for further details.
Transfer learning and ensemble learning. To ensure sufficient patient cases, we first trained LightGBM and MLP models in MIMIC and later transferred them to the RJ dataset in Task 2. These models were selected based on the performance of Task 1 and were used as representatives of traditional machine learning models and neural network models. They needed to be retrained from Task 1 because some variables were not available in the RJ database. Due to the population differences between the MIMIC and RJ databases, the data were standardized separately as in Task 1. Additionally, similar features (the maximum, average, median, and minimum values of variables with low recording frequency at the same hour) are reduced to one feature to reduce the dimension and help the transfer. The parameters of each model were shared as initial parameters and tuned again during training on RJ database.
Finally, we integrated the MLP and LightGBM models by taking the average to make the results

more robust and accurate. This result was used as the final output of the sepsis prediction model.

Role of the funding source

The funders of the study had no role in the study design, data collection, data analysis, data interpretation, or writing of the report.

Results

Sepsis prediction model

We evaluated our models based on accuracy, the area under the receiver operating characteristic curve (AUC), sensitivity, and specificity of the test set.

Performance on the MIMIC-III dataset

In Task 1, the AUCs of XGBoost and LightGBM were the highest among these sepsis prediction models, followed by MLP and LSTM, and SVM performed the worst. Appendix VII shows the full performance of the five models. Although the GBM structure is relatively simple, it outperforms artificial neural network models. We compared our models with other models trained on the same open-source database, MIMIC, using the Sepsis-3 criteria, and reported prediction results within 5 h before the onset of sepsis, including InSight,[8] AISE,[9] MGP-TCN, DTW-KNN,[10] MLA,[11] DSPA,[12] and MGP-AttTCN.[16] Table 1 shows that our models basically outperform the others. However, it should be highlighted that although these models all use the MIMIC-III database, there were still differences in the training and test sets due to differences in the details of case extraction and sepsis criteria; therefore, the comparison is not as standard as most machine learning comparison works based on benchmark data.

Table 1 The results of different models on the MIMIC-III dataset

	Preceding hours	Accuracy	AUC	Sensitivity	Specificity
InSight[8]	4	0·57	0·74	0·80	0·54
AISE[9]	4	0·64	0·84	0·85	0·64
MGP-TCN[10]	4	-	about 0·85	-	-
DTW-KNN[10]	4	-	about 0·88	-	-
MLA[11]	0	-	0·88	0·80	0·78
MLA[11]	24	-	0·84	0·80	0·72
DSPA[12]	4	-	0·982	-	-
MGP-AttTCN[16]	4	-	0·746	-	-
LightGBM	4	0·9075	0·9833	0·8491	0·9660
MLP	4	0·8462	0·9564	0·7292	0·9632

Performance on the RJ dataset

In Task 2, after transfer learning and ensemble learning, the final performance of the sepsis prediction model is shown in Table 2. The overall performance was similar to that of LightGBM or MLP transferred models, and the average value of AUC was slightly higher than that of LightGBM or MLP. Detailed results of transfer learning can be found in Appendix VIII.

Table 2 The results of ensemble model in Task 2

Preceding hours	Accuracy	AUC	Sensitivity	Specificity
1	0·8250	0·9231	0·6667	0·9833

2	0·7944	0·9200	0·6000	0·9889
3	0·8417	0·9242	0·7083	0·9750
4	0·8467	0·9436	0·6933	1·0000
5	0·8639	0·9065	0·7389	0·9889

Feature interpretability

We used Shapley additive explanation (SHAP)[24] analysis to explore the importance of the features. For LightGBM models in SEPRES, antibiotics, respiratory rate, total positive end-expiratory pressure level, fibrinogen level, temperature, net balance, and age were important in most models. For MLP models, age, respiratory rate, ventilation, heart rate, antibiotics, and temperature were important in most models. Some of these features (antibiotics, respiratory rate, temperature, ventilation, and heart rate) were related to the definition of Sepsis-3 or SIRS, while there is also literature arguing for an association between some of these features (respiratory rate,[25] fibrinogen,[26] net balance,[27] and age[28]) and the severity or mortality of sepsis. Detailed results are provided in Appendix IX.

SEPRES

Model inference in SEPRES

The detailed steps of model inference are as follows:

1) Obtain real-time features of patients using SQL query statements.
2) The features were standardized by calling the scaler obtained in the training set.
3) Call the trained model to get the prediction results.
4) Call the interpretive tool to get the importance of the features based on the prediction results.
5) Output and store prediction results and interpretations in a standard format.

System operation

SEPRES provides predictions and explanations for every patient in the ICU every hour, including the risk of sepsis onset in the next 5 h, the influence of features on the predictions calculated by SHAP, and SOFA predictions. It has been operating at Ruijin Hospital for several months, providing hourly early warning services for over 100 patients in the ICU.

The PC terminal of the user interface is a layout in the ICU common room. Figure 3 presents an example of the display board for all patients in the ICU, including predictions of sepsis onset and SOFA changes, where high and low risks are shown by red and blue bars, respectively. Figure 4 shows the data details for a specific patient to observe the current and past status of the patient.

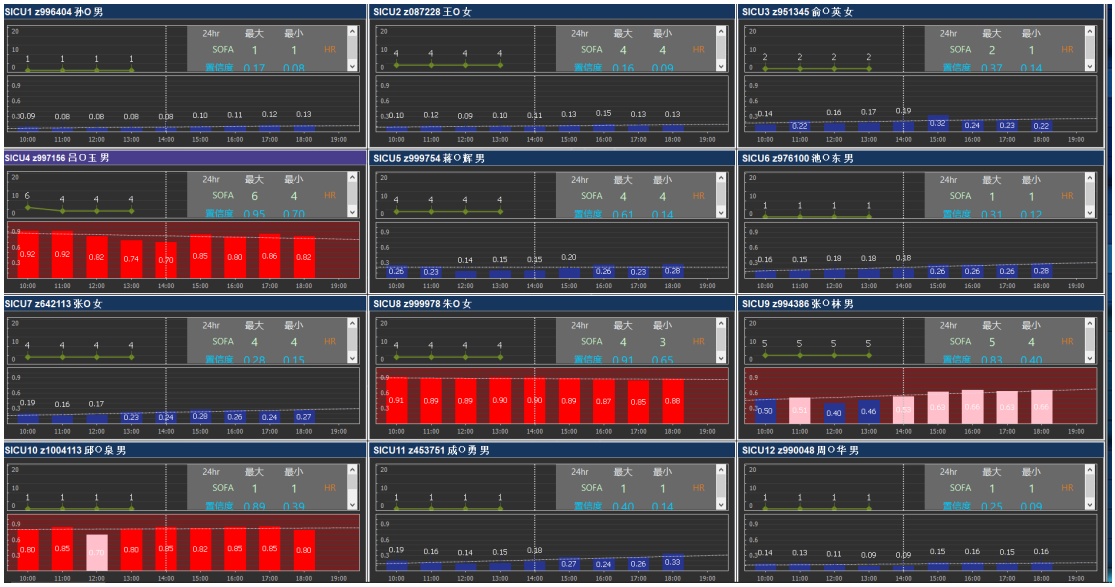


Figure 3 The user interface example. Each subplot shows the change in SOFA score and sepsis-onset prediction for a patient. It is subtitled with patient information, and the upper right corner shows the maximum and minimum Sequential Organ Failure Assessment scores and sepsis-onset predictions within 24 h.

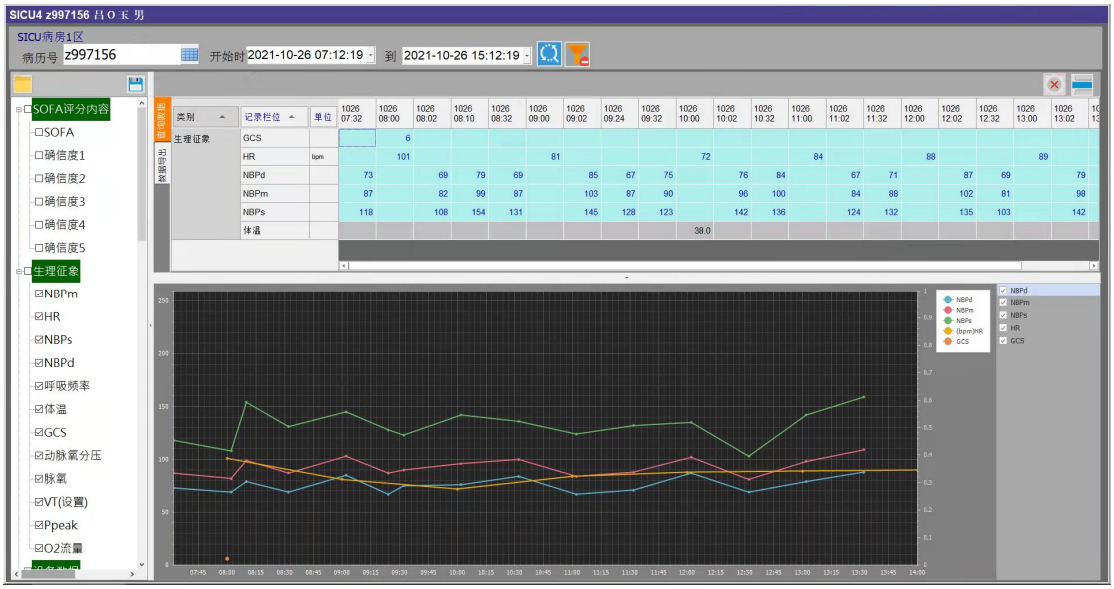


Figure 4 Historical data review for individual patients. The title is patient information, below the title are filter criteria, on the left side are optional data types, and the rest is the table and chart for the selected data.

Real-time performance of the sepsis prediction model in the ICU of the Ruijin Hospital

We extracted a total of 67 ICU stays from February 2021 to June 2021 from the system. Each stay was labeled by the change in SOFA score and the doctor's examination for infection (based on antibiotics or blood culture), and 40 of these stays were labeled as having at least one sepsis onset at a threshold of 0·5. Data from the control group and near onset of sepsis in the case group were included in the analysis. The statistical results of the predictions are presented in Table 3.

Table 3 The results of real-time data

Preceding hours	Accuracy	AUC	Sensitivity	Specificity
1	0·8243	0·8636	0·8281	0·8150
2	0·8448	0·8789	0·8682	0·7843
3	0·8483	0·8992	0·8626	0·8106
4	0·8533	0·8952	0·8802	0·7803
5	0·8583	0·8858	0·8951	0·7566

Case studies

We discussed several of these cases, including true-positive, true-negative, false-positive, and false-negative cases. Figure 5 illustrates the model predictions for positive cases near the onset of sepsis or negative cases over a random period (see Appendix X for more details).

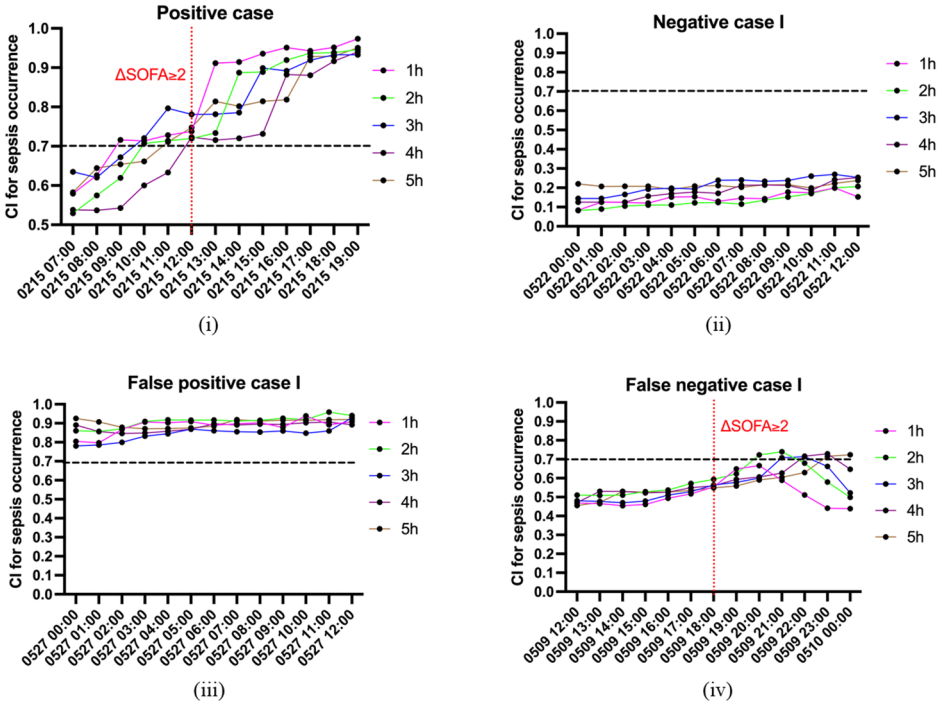


Figure 5 some illustrative examples of the prediction. Each subplot describes the confidence index (CI) for multiple models (Y-axis) at the target time (X-axis). (i) The condition of the patient aggravated in the early morning of February 15, 2021, with multiple organ dysfunction, and the patient was diagnosed with sepsis at 12:00 AM. Our model prediction exceeded the warning threshold of 0·7 for the prediction at 9:00 AM. (ii) Despite the high Sequential Organ Failure Assessment (SOFA) score (7·0), there was no evidence of ΔSOFA $\geqslant 2$ within 72 h. Consistently, the predictions were all lower than the threshold. (iii) Although the patient's SOFA score was stable at 6·0, our model made incorrect predictions. (iv) The SOFA score showed an increase from 6·0 to 9·0 at 06:00 PM on May 9, 2021. In combination with evidence of infection, the patient was diagnosed with sepsis. However, the prediction was below the warning threshold.

Our model can detect sepsis early in most cases, although there are a small number of false-negative and false-positive cases. We propose that the possible reason for false-negatives is that our models tend to give lower predictions when the collected data are limited. The early prediction of sepsis occurrence by our model effectively guided medical practitioners to pay

more attention appropriately to this patient, leading to early diagnosis of sepsis and more efficient management of ICU patients.

Discussion

Machine learning methods have been considered a promising method for early warning of sepsis in the ICU.[6-17] Early diagnosis and timely management of septic patients can effectively improve prognosis.[29] However, sepsis may not be diagnosed in time in the clinic due to the doctor's shift and day-night shift of the medical staff. Therefore, an accurate and efficient early prediction system for sepsis at the bedside is important.

In this study, we established an ICU bedside sepsis early warning system, SEPRES, to conduct real-time sepsis prediction for patients in the ICU through a data integration system. In contrast to most studies on machine-learning sepsis prediction in the open-source database, SEPRES has been developed and conducted real-time inference and analysis in the ICU of the Ruijin Hospital by the integration of IntelliVue Information Center, Ventilators, Philips ICCA system, LIS, and HIS data. Additionally, the system can display the patient's historical data in the user interface to help doctors intuitively obtain changes in the patient's condition. Although SEPRES could not provide a definitive basis for our therapeutic regime, the probability of sepsis occurrence allows us to pay more attention to specific patients. Furthermore, weight analysis of medical factors can provide insights into the use of therapeutic regimens.

To avoid the influence of the insufficiency of data size and inhomogeneity distribution of the data in different medical centers on training and inferring machine learning models, we deployed the transfer learning technique to improve the performance of the specific medical center. In particular, MIMIC-III mainly enrolled white patients (40996 of 58976 hospital admissions), in contrast to the private dataset of the Ruijin Hospital that is mainly composed of the Chinese population, which has a significant difference in certain features of the sepsis prediction model (See Appendix XI). In this way, the transfer learning process improved the prediction AUC of LightGBM models from 0·8613–0·8913 to 0·8964–0·9348 on the historical data of the Ruijin Hospital.

The interpretive tool may help medical practitioners identify risk factors. In the SHAP analysis of the models loaded in SEPRES, we paid special attention to the insights brought about by the importance of net balance. As shown in Supplementary Figure 2, a positive net balance indicates a higher risk of sepsis. Because the net balance is nursing data that are difficult to collect, in the MIMIC dataset, 4079 out of 6891 episodes have no balance data for colloid and crystalloid solutions, while in the RJ dataset, only 25 out of 329 episodes had no net balance data. Therefore, net balance has not been considered a feature for most machine learning models or has been analyzed as an important factor for sepsis prediction inference based on MIMIC datasets. Our SHAP analysis showed that a negative net balance tended to decrease the predicted probability of sepsis. Indeed, positive cumulative fluid balance has been reported to be an independent predictor of ICU mortality.[27] Furthermore, Lin et al. have shown that patients with an early positive fluid balance have an increased risk of developing venous thromboembolism.[30] The weight of net balance, as shown in our SHAP analysis, further emphasized the importance of careful fluid management in critically ill patients. Therefore, we argue that including the net balance in the prediction model may improve not only the performance but also the clinical management in the ICU.

Our model has certain limitations. First, we enrolled only patients who were non-septic during the entire period in the ICU as negative controls. The enrollment condition may be too pure to be used to establish a model to predict the onset period of sepsis, which may cause false-positive cases. Second, as we observed in consecutive case studies, patients diagnosed with sepsis shortly after being transferred to the ICU were difficult to predict by the model. That is, a short period of data recording may cause false-negative cases. Finally, our model incorporates variables, such as antibiotics and mechanical ventilation, resulting in the predictions of the model being influenced by the subjective behavior of the doctor.

These limitations will be addressed in future work through diverse methods, including fine-grained labeling, inclusion of data collection from the ICU, and data augmentation.

Moreover, this workflow applies to disease warnings other than sepsis in the ICU, such as disseminated intravascular coagulation and acute kidney injury, with the help of the data integration system to collect the necessary features and data for model construction.

Contributors

WL, YT, LL, QC, RL, and Lin C conceived and designed the study. YT, LL, QC, RL, DC, HQ, and YH acquired the data. WL, Lin C, and Lai C implemented quality control of data and algorithms. WL, QC, RL, and Lai C had full access to and verified all data in the study. QC developed, trained, and applied machine-learning models. Lai C developed a data integration system. YT, LL, RL, DC, and HQ performed the consecutive case studies. QC and RL prepared the first draft of the manuscript. WL, LL, and YT revised the manuscript. All authors contributed to the preparation of the manuscript.

Declaration of Interest

We report no competing interests.

Data Sharing

The MIMIC-III database can be accessed at https://physionet.org/content/mimiciii/1.4/ after becoming a member of Physionet (https://physionet.org/). The RJ database used in this study is not publicly available. The code used to develop the model in this manuscript is available from the corresponding author upon reasonable request.

Acknowledgments

The authors acknowledge Shanghai Electric Group Co., Ltd. Central Academe for their support during the development of the data integration system.

Ethics Committee Approval

Our study was approved by the Ruijin Hospital Ethics Committee [ethics committee reference number: (2020) Linlunshen No. (140)].

References

1 Cecconi M, Evans L, Levy M, Rhodes A. Sepsis and septic shock. *Lancet* 2018; **392**: 75–87.

2 Singer M, Deutschman CS, Seymour CW, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). *Jama* 2016; **315**: 801–10.

3 Angus DC, Linde-Zwirble WT, Lidicker J, Clermont G, Carcillo J, Pinsky MR. Epidemiology of severe sepsis in the United States: Analysis of incidence, outcome, and associated costs of care. *Crit Care Med* 2001; **29**: 1303–10.

4 Marik PE, Farkas JD. The changing paradigm of sepsis: early diagnosis, early antibiotics, early pressors, and early adjuvant treatment. *Crit Care Med* 2018; **46**: 1690–2.

5 Filbin MR, Lynch J, Gillingham TD, et al. Presenting symptoms independently predict mortality in septic shock: Importance of a previously unmeasured confounder. *Crit Care Med* 2018; **46**: 1592–9.

6 Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. *Sci Transl Med* 2015; **7**: 299ra122–299ra122.

7 Lauritsen SM, Kalør ME, Kongsgaard EL, et al. Early detection of sepsis utilizing deep learning on electronic health record event sequences. *Artif Intell Med* 2020; **104**: 101820.

8 Desautels T, Calvert J, Hoffman J, et al. Prediction of sepsis in the intensive care unit with minimal electronic health record data: A machine learning approach. *JMIR Med Inform* 2016; **4**: e5909.

9 Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An interpretable machine learning model for accurate prediction of sepsis in the ICU. *Crit Care Med* 2018; **46**: 547.

10 Moor M, Horn M, Rieck B, Roqueiro D, Borgwardt K. Early recognition of sepsis with Gaussian process temporal convolutional networks and dynamic time warping. Machine Learning for Healthcare Conference; 2019: PMLR.

11 Barton C, Chettipally U, Zhou Y, et al. Evaluation of a machine learning algorithm for up to 48-hour advance prediction of sepsis using six vital signs. *Comput Biol Med* 2019; **109**: 79–84.

12 Asuroglu T, Ogul H. A deep learning approach for sepsis monitoring via severity score estimation. *Comput Methods Programs Biomed* 2021; **198**: 105816.

13 McCoy A, Das R. Reducing patient mortality, length of stay and readmissions through machine learning-based sepsis prediction in the emergency department, intensive care unit and hospital floor units. *BMJ Open Qual* 2017; **6**: e000158.

14 Shimabukuro DW, Barton CW, Feldman MD, Mataraso SJ, Das R. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: A randomised clinical trial. *BMJ Open Respir Res* 2017; **4**: e000234.

15 Zhang D, Yin C, Hunold KM, Jiang X, Caterino JM, Zhang P. An interpretable deep-learning model for early prediction of sepsis in the emergency department. *Patterns (N Y)* 2021; **2**: 100196.

16 Rosnati M, Fortuin V. MGP-AttTCN: An interpretable machine learning model for the prediction of sepsis. *PLoS One* 2021; **16**: e0251248.

17 Oei SP, van Sloun RJ, van der Ven M, Korsten HH, Mischi M. Towards early sepsis detection from measurements at the general ward through deep learning. *Intell Based Med*

2021; **5**: 100042.
18 Johnson AE, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016; **3**: 1–9.
19 Sorani MD, Hemphill JC, Morabito D, Rosenthal G, Manley GT. New approaches to physiological informatics in neurocritical care. *Neurocrit Care* 2007; **7**: 45–52.
20 Goldstein B, McNames J, McDonald BA, et al. Physiologic data acquisition system and database for the study of disease dynamics in the intensive care unit. *Crit Care Med* 2003; **31**: 433–41.
21 Sun Y, Guo F, Kaffashi F, Jacono FJ, DeGeorgia M, Loparo KA. INSMA: An integrated system for multimodal data acquisition and analysis in the intensive care unit. *J Biomed Inform* 2020; **106**: 103434.
22 Chen T, Guestrin C. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016: 785–94.
23 Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems 2017; **30**: 3146–54.
24 Lundberg S, Lee SI. A Unified approach to interpreting model predictions. Proceedings of the 31st international conference on neural information processing systems 2017: 4768–77.
25 Kenzaka T, Okayama M, Kuroki S, et al. Importance of vital signs to the early diagnosis and severity of sepsis: Association between vital signs and sequential organ failure assessment score in patients with sepsis. *Intern Med* 2012; **51**: 871–6.
26 Matsubara T, Yamakawa K, Umemura Y, et al. Significance of plasma fibrinogen level and antithrombin activity in sepsis: A multicenter cohort study using a cubic spline model. *Thromb Res* 2019; **181**: 17–23.
27 Brotfain E, Koyfman L, Toledano R, et al. Positive fluid balance as a major predictor of clinical outcome of patients with sepsis/septic shock after ICU discharge. *Am J Emerg Med* 2016; **34**: 2122–6.
28 Martin GS, Mannino DM, Moss M. The effect of age on the development and outcome of adult sepsis. *Crit Care Med* 2006; **34**: 15–21.
29 Burdick H, Pino E, Gabel-Comeau D, et al. Effect of a sepsis prediction algorithm on patient mortality, length of stay and readmission: A prospective multicentre clinical outcomes evaluation of real-world patient data from US hospitals. *BMJ Health Care Inform* 2020; **27**: e100109.
30 Lin T-L, Dhillon NK, Conde G, et al. Early positive fluid balance is predictive for venous thromboembolism in critically ill surgical patients. *Am J Surg* 2021; **222**: 220–6.

Research in context

Evidence before this study

We searched PubMed from inception to September 30, 2021, using the keywords “machine learning” or “deep learning” or “artificial intelligence” and “sepsis,” and using the keywords “critical care” or “ICU” or “sepsis” and “data integration system” or “data acquisition system” or “integrated,” without language restrictions. Previous studies built various prediction models based on different sepsis definitions and datasets or built data acquisition systems to integrate some types of data. However, most of them did not combine the two together to get a real-time prediction and describe the whole process in detail.

Added value of this study

We developed an entire sepsis prediction system in the ICU, including a set of procedures that can be applied in practice. The machine learning models achieved high AUC scores in the two databases, and we interpreted the predictions. In addition, we examined our system through consecutive case studies. Moreover, this workflow is also applicable to other disease warnings, not only for sepsis.

Implications of all the available evidence

Our system can be applied to display patients’ conditions in real-time, identify patients more likely to suffer from sepsis, help medical practitioners focus on them, and help with future research.

supplementary material

Contents

Appendix I Details of data extraction 1

Appendix II Details of feature extraction 2

Appendix III Complete list of variables used in Task 1 2

Appendix IV Complete list of variables used in Task 2 3

Appendix V Machine-learning models 3

Appendix VI Training method 4

Appendix VII Results of sepsis prediction models in Task 1 5

Appendix VIII Results of transferred models in Task 2 6

Appendix IX Feature importance 7

Appendix X Case studies 9

Appendix XI Differences in features between MIMIC dataset and RJ dataset 15

Reference 16

Appendix I Details of data extraction

Serial Port

Most medical devices communicate through the network or the serial port, including RS-232 and mini-DIN. The serial port uses binary signals, so the data rate in bits per second is equal to the symbol rate in baud, and commonly supported bit rates include from 2400 to 115200 bits per second.

The HL7 interface

The device data are transmitted to the data integration system through the HL7 interface, an electronic data interchange standard for the provision of inpatient care based on the IP protocol. Using TCP/IP connections, client systems can obtain data from the interface using both active sends and queries. HL7 v2.3.1 is generally used.

Supplementary Table 1 The types of data collected from the devices and systems

Source device/system	Data type	Output medium	Format
IntelliVue Information Center	Vital signs	Network	HL7
PB 840 Ventilator Maquet Servo-i Ventilator Maquet Servo-s Ventilator	Ventilator data	RS-232	According to the device output format
Philips ICCA	Pharmacy data, GCS, and urine output	Network	WebServices
Laboratory Information System	Laboratory data	Network	WebServices
Hospital Information System	Admission, discharge, and hospitalization data	Network	WebServices

Appendix II Details of feature extraction

Seventy-eight patient variables from MIMIC were chosen as raw data for the dataset. Appendix III contains a complete list of these variables. We excluded significantly incorrect records by setting the range of variables according to the specialists. When integrating the same variables from different sources, we set priorities to extract values with the highest confidence. After data cleaning, these data were summarized per hour into the maximum, average, median, and minimum, except for some changeless or durative variables, which in total were 285 features. Padding was used if there was no value at the corresponding time. Padding values were taken as the nearest value before, or the average of all patients when no value was valid since the patient's admission. Episodes with too few valid variables were removed to ensure data quality. We used a five-hour-long time window from the episodes to predict sepsis; thus, each sample point in these tasks had 1425 features. Finally, we obtained a dataset with 1057 positive and 5834 negative episodes. We divided the dataset into training, validation, and test sets. Negative episodes were divided at a ratio of 7:1:2. For positive episodes, we chose the same number as the negative episodes in the validation and test sets. The remaining positive episodes were included in the training set. Oversampling of positive sample points or down-sampling of negative sample points was used to ensure that the proportion was 1:1 in each set.

Similar preparation steps were used in the RJ database as well, although only sixty-three variables were available. Appendix IV contains a complete list of these variables. These variables were summarized hourly into 226 features. After padding and filtering, we obtained the RJ dataset with 115 positive and 239 negative episodes. We divided them into a training set with 76 positive and 200 negative episodes, a validation set with nine positive and nine negative episodes, and a test set with 30 positive and 30 negative episodes.

Appendix III Complete list of variables used in Task 1

We consulted literature on predicting sepsis or SOFA scores and the variables mentioned in the literature and could be extracted from MIMIC-III were selected, for a total of 78 variables.[1-4]

The 78 variables are MAP, heart rate, O_2sat, SBP, DBP, respiratory rate, temperature, GCS, PaO_2, FiO_2, SpO_2, cardiac output, stroke volume, stroke volume variation, tidal volume, peak inspiratory pressure, total PEEP level, O_2 flow rate, WBC, hemoglobin, hematocrit, creatinine, bilirubin, bilirubin direct, platelets, INR, PTT, AST, lactate, glucose, potassium, calcium, BUN, phosphorus, magnesium, chloride, BNP, troponin I, fibrinogen, CRP, sedimentation rate, ammonia, PH, PCO_2, bicarbonate, base excess, SaO_2, anion gap, albumin, bands, PT, sodium, ferritin, transferrin, creatine kinase, creatine kinase-MB, LDH, troponin T, RDW, ALP, MCHC, uric acid, monocytes, lymphocytes, MCH, $AaDO_2$, RBC, MCV, neutrophils, weight, urine output in the past 24 hours, net balance, ventilation, number of antibiotics in the past 12, 24, and 48 hours, SOFA, and age.

Appendix IV Complete list of variables used in Task 2

Based on Appendix II, we removed some variables that were not recorded in the Ruijin Hospital historical database and removed some infrequently used variables based on doctors' opinions. We also added four variables that were involved in the SOFA score.

Nineteen variables were removed: GCS, O_2sat, Cardiac Output, Stroke Volume, Stroke Volume Variation, Calcium, BNP, CRP, Sedimentation Rate, Ammonia, Anion Gap, Bands, Ferritin, Transferrin, Troponin T, RDW, MCHC, MCH, MCV.

Four variables were added: rate of norepinephrine, epinephrine, dopamine, and dobutamine.

Finally, 63 variables were collected.

Appendix V Machine-learning models

1. Support Vector Machine (SVM)

The rationale of Support Vector Machine (SVM) is to find such a hyperplane $y = \boldsymbol{w} \cdot \boldsymbol{x} + b$ to separate data that the distance between positive data on one side and the hyperplane and the distance between negative data on the other side and the hyperplane is maximized:

$$\max_{\boldsymbol{w},b} \quad \gamma$$
$$s.t. \quad y^n \left(\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \cdot \mathbf{x}^n + \frac{b}{\|\boldsymbol{w}\|} \right) \geq \gamma, \ \ \forall n = 1, 2, \cdots, N$$

Where the dataset is $\left\{ \left(\mathbf{x}^n, y^n \right) \right\}_{n=1}^{N}$, and y^n is the label of $\mathbf{x}^n$.

Furthermore, kernel tricks can be introduced to improve the nonlinearity of the model. Data are mapped into a feature space using a nonlinear mapping with the help of kernel function, including polynomial kernel function, linear kernel, or mixed kernel function.

2. Multi-Layer Perceptron (MLP)

Multi-Layer Perceptron (MLP) is a classical and useful neural network model. It contains an input layer, several hidden layers, and an output layer.

Let *In* denote the input layer, and $Fc_1, Fc_2, \ldots, Fc_L$ denote hidden layers and the output layer successively, where Fc_L is the output layer. Each hidden layer and output layer learn a nonlinear map $\mathbf{h}_i = \sigma\left(\mathbf{W}_i \mathbf{h}_{i-1} + \mathbf{b}_i \right)$, where $\mathbf{h}_i$ is the representation of an example at layer Fc_i, $\mathbf{W}_i$ and $\mathbf{b}_i$ are the weight and bias parameters of layer Fc_i, and σ is the activation function, taken as Rectifier Linear Units (ReLu) $\sigma(\mathbf{x}) = max(0, \mathbf{x})$ for the hidden layers and Softmax Units $\sigma(\mathbf{x})_i = exp(\mathbf{x}_i) / \sum_j exp(\mathbf{x}_j)$ for the output layer. The loss function like

$$L(f(\mathbf{x}^n), y^n)$$
$$= -\frac{1}{N}\sum_{n=1}^{N}\left[y^n \log\left(f\left(\mathbf{x}^n\right)\right) - \left(1-y^n\right)\log\left(1-f\left(\mathbf{x}^n\right)\right)\right] + \lambda\sum_{i=1}^{L}\|\mathbf{W}_i\|_F$$

and Stochastic Gradient Descent algorithms are always used to train the networks, where $f\left(\mathbf{x}^n\right)$ is the network output of $\mathbf{x}^n$, and $\|\cdot\|_F$ is the Frobenius norm. The first term in the formula is the cross-entropy loss and the second is a weight decay term in order to prevent over-fitting where λ is the weight decay coefficient.

3. Gradient Boosting Decision Tree

Gradient boosting decision tree (GBDT) is a widely-used machine learning algorithm and there are many effective implementations such as XGBoost and LightGBM which have been used in our experiments. XGBoost makes use of second-order information of loss function instead of first-order information in the case of GBDT, which makes convergence faster and better. Moreover, XGBoost uses several algorithms and technologies to accelerate the training procedure. Compared with XGBoost, LightGBM can accelerate the training further, benefiting from the histogram-based algorithm, gradient-based one-side sampling, and exclusive feature bundling.

4. Long Short-Term Memory (LSTM)

Considering the time-sequential datasets, recurrent neural networks (RNN), specifically long short-term memory (LSTM) networks, which is the most successful RNN, had been used in our experiments. LSTM networks have memory blocks consisting of memory cells and gates in the recurrent hidden layer. The gates control which information to forget and which to remember. Through these, LSTM networks are able to store information selectively and capture the long-time-dependent behavior compared with MLP or GBDT. A single repeating structure of the LSTM can be represented as the following:

$$\begin{aligned}
\boldsymbol{f}_t &= \sigma\left(\boldsymbol{W}_f \cdot \left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{b}_f\right) \\
\boldsymbol{i}_t &= \sigma\left(\boldsymbol{W}_i \cdot \left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{b}_i\right) \\
\tilde{\boldsymbol{C}}_t &= \tanh\left(\boldsymbol{W}_C \cdot \left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{b}_C\right) \\
\boldsymbol{C}_t &= \boldsymbol{f}_t * \boldsymbol{C}_{t-1} + \boldsymbol{i}_t * \tilde{\boldsymbol{C}}_t \\
\boldsymbol{o}_t &= \sigma\left(\boldsymbol{W}_o \cdot \left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{b}_o\right) \\
\boldsymbol{h}_t &= \boldsymbol{o}_t * \tanh\left(\boldsymbol{C}_t\right)
\end{aligned}$$

Where $\boldsymbol{f}_t, \boldsymbol{i}_t, \boldsymbol{C}_t, \boldsymbol{o}_t, \boldsymbol{h}_t$ represent the forget gate, input gate, cell state, output gate, hidden state, and the symbol $*$ refers to element-wise multiplication.

Appendix VI Training method

In Task 1, some redundant features were removed to accelerate the training of the SVM and MLP. For SVM, all average features and features whose coefficient of variation was > 2 were used; for MLP, only one of the maximum, average, median, and minimum of each laboratory variable was kept, considering

the low record frequency. Data were standardized (i.e., each feature's value range was linearly scaled between 0 and 1) before training to eliminate magnitude differences among features and reduce distribution differences between the two datasets in the next task.

We chose the linear kernel function and 1 as the penalty factor in the SVM. For MLP, 460 features were selected as the input. We finally used a six-layer architecture, which is shown in Supplementary Table 2. We chose 256 as the batch size, 0·001 as the learning rate, 0·6 as the dropout rate, and 0·001 as the weight decay coefficient in MLP. We also used dropout to prevent over-fitting as well. The proposed networks were trained with the AdaGrad algorithm. For XGBoost, we set max_depth as 6, colsample_bytree as 0·2, and other configurations as default. For LightGBM, we set num_leaves as 5, lambda_l2 as 0·1, learning rate as 0·2, and other configurations as default. During LSTM training, we used LSTM architecture with four hidden LSTM layers with 16 one-cell memory blocks and a fully connected layer with one output unit added, followed by a sigmoid function. We set the learning rate as 0·0001 and the batch size to 2000.

Supplementary Table 2 The network structure of the MLP

	input	layer1	layer2	layer3	layer4	layer5	output
units	460	256	128	64	24	24	2

Appendix VII Results of sepsis prediction models in Task 1

Supplementary Table 3 THE RESULTS OF MODELS IN TASK 1

	Preceding hours	Accuracy	AUC	Sensitivity	Specificity
SVM	1	0·8762	0·8762	0·8726	0·8797
	2	0·8687	0·8687	0·8711	0·8664
	3	0·8791	0·8791	0·8762	0·8821
	4	0·8665	0·8665	0·8736	0·8594
	5	0·8719	0·8719	0·8711	0·8726
MLP	1	0·8443	0·9539	0·7123	0·9764
	2	0·8412	0·9546	0·7107	0·9717
	3	0·8414	0·9579	0·7123	0·9705
	4	0·8462	0·9564	0·7292	0·9632
	5	0·8534	0·9552	0·7437	0·9631
XGBoost	1	0·8903	0·9867	0·7972	0·9835
	2	0·8962	0·9883	0·8113	0·9811
	3	0·8950	0·9875	0·8149	0·9752
	4	0·8976	0·9872	0·8189	0·9764
	5	0·8990	0·9873	0·8200	0·9780
LightGBM	1	0·8892	0·9839	0·8090	0·9693
	2	0·8954	0·9840	0·8192	0·9717
	3	0·8950	0·9825	0·8255	0·9646
	4	0·9075	0·9833	0·8491	0·9660

	5	0·9076	0·9848	0·8420	0·9733
LSTM	1	0·8384	0·9463	0·7028	0·9741
	2	0·8522	0·9541	0·7421	0·9623
	3	0·8561	0·9486	0·7665	0·9458
	4	0·8509	0·9492	0·7642	0·9377
	5	0·8671	0·9485	0·8176	0·9167

Supplementary Table 3 shows that the AUCs of XGBoost and LightGBM were the highest among these sepsis prediction models, followed by MLP and LSTM, and SVM performed the worst. Although the GBM's structure was relatively simple, it outperformed the artificial neural network models, which may be due to the complexity of artificial neural network models, leading to poor generalization. At the same time, LSTM didn't outperform MLP. In each model, the five tasks of predicting sepsis from 1 to 5 h in advance did not show significant differences in the predicting AUC.

Appendix VIII Results of transferred models in Task 2

In Task 2, the performances of LightGBM and MLP models trained directly on the RJ dataset are shown in the top half of Supplementary Table 4 and Supplementary Table 5 as a comparison, while models trained with the transfer learning technique are shown in the bottom half. Models trained using the transfer learning technique performed better in most cases. This indicates that the models can learn generic knowledge from MIMIC and help with predictions on the RJ dataset. These models were further ensembled by taking the average.

Supplementary Table 4 THE RESULTS OF LIGHTGBM IN TASK 2

Preceding hours	Transfer learning	Accuracy	AUC	Sensitivity	Specificity
1	N	0·8500	0·8644	0·7333	0·9667
2	N	0·7333	0·8769	0·5889	0·8778
3	N	0·7292	0·8613	0·5833	0·8750
4	N	0·7600	0·8857	0·5733	0·9467
5	N	0·7556	0·8913	0·5889	0·9222
1	Y	0·8167	0·9183	0·7333	0·9000
2	Y	0·7833	0·9348	0·5667	1·0000
3	Y	0·8583	0·9171	0·7333	0·9833
4	Y	0·7900	0·9311	0·5800	1·0000
5	Y	0·8472	0·8964	0·7056	0·9889

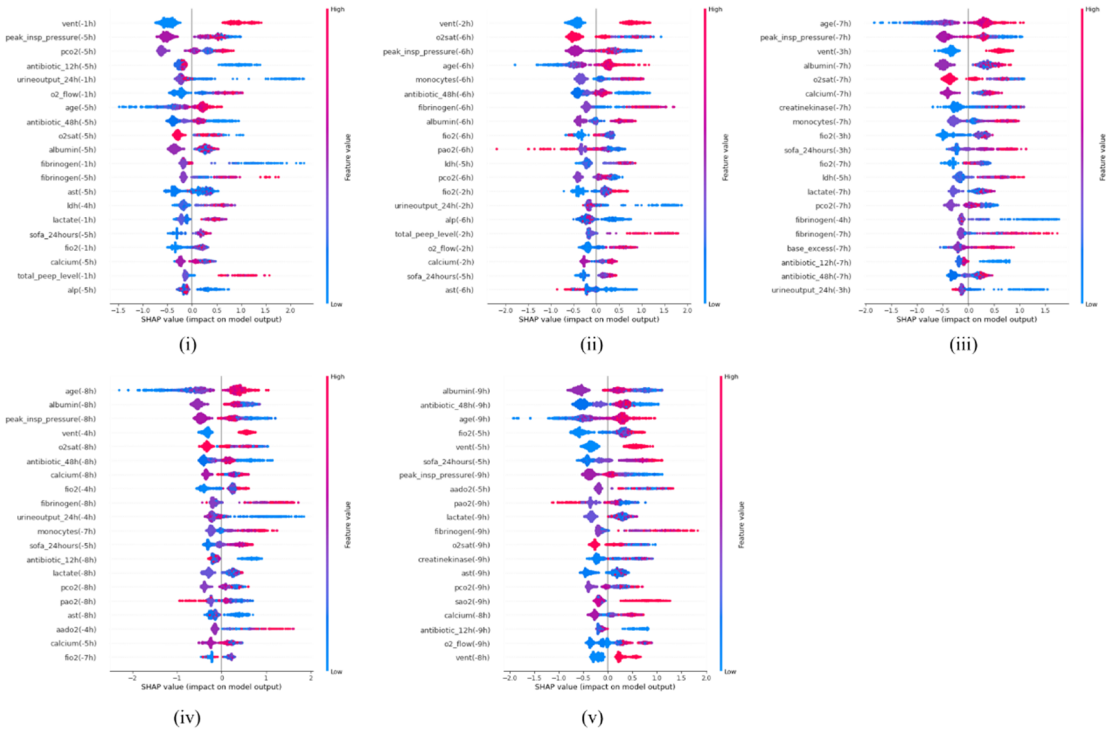
Supplementary Table 5 THE RESULTS OF MLP IN TASK 2

Preceding hours	Transfer learning	Accuracy	AUC	Sensitivity	Specificity
1	N	0·7750	0·9150	0·5667	0·9833
2	N	0·7778	0·9037	0·5889	0·9667
3	N	0·8083	0·9021	0·6333	0·9833

Preceding hours	Transfer learning	Accuracy	AUC	Sensitivity	Specificity
4	N	0·7900	0·8953	0·6133	0·9667
5	N	0·8000	0·9010	0·6222	0·9778
1	Y	0·8083	0·9172	0·6500	0·9667
2	Y	0·7944	0·8936	0·6000	0·9889
3	Y	0·8333	0·9166	0·6917	0·9750
4	Y	0·8267	0·9221	0·6867	0·9667
5	Y	0·8250	0·9208	0·6778	0·9722

Appendix IX Feature importance

We used the Shapley Additive Explanations (SHAP) analysis to explore the importance of features of different models. The SHAP value of each feature represents the impact of the feature on the output of the model. For LightGBM models trained on the MIMIC-III dataset, the beeswarm plots were shown in Supplementary Figure 1. The importance of features of the same hour and from the same variable were combined, and the twenty most important features were displayed in the figures. As shown in Supplementary Figure 1, several features were found to be important in most models, e.g., FiO_2, fibrinogen, calcium, ventilation, age, albumin, antibiotics, O_2Sat, PCO_2, peak inspiratory pressure, and SOFA.

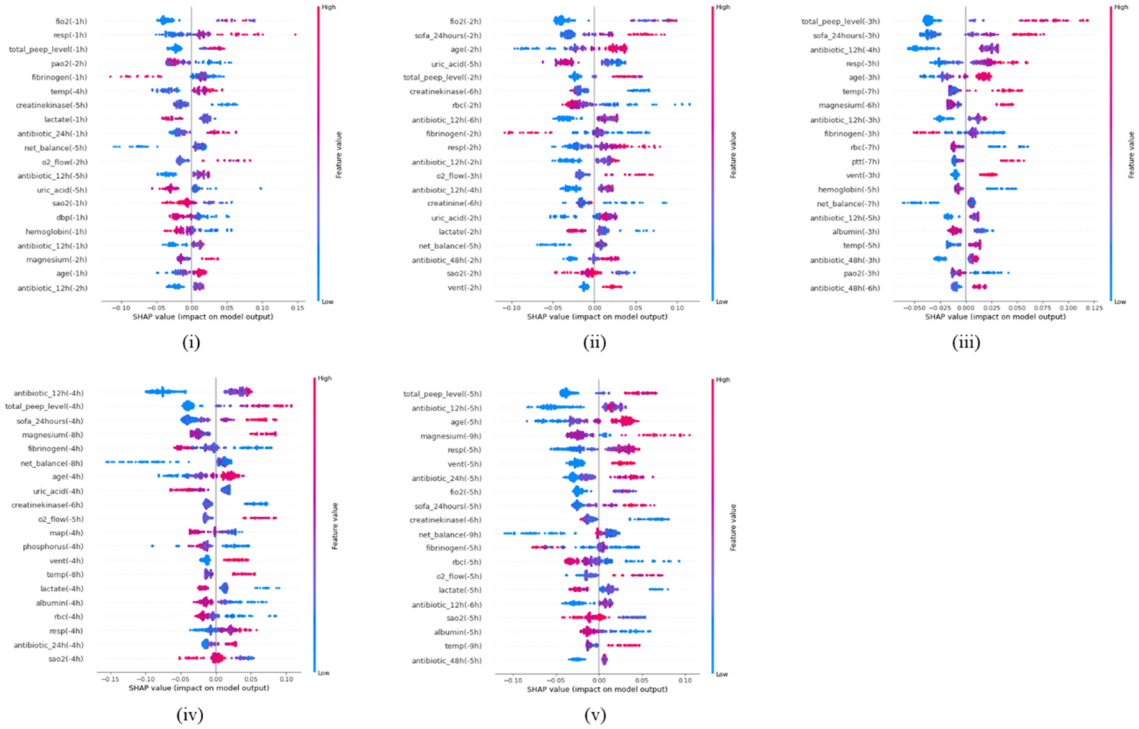


Supplementary Figure 1 Feature importance of LightGBM models on the MIMIC-III dataset. (i) - (v) are the beeswarm plots of the models detecting sepsis in 1 - 5 hours preceding, respectively. In each subgraph, Y-axis represents different features and X-axis represents the SHAP value of the sample point represented by one dot. A vertical thickness of a dot cluster represents the number of

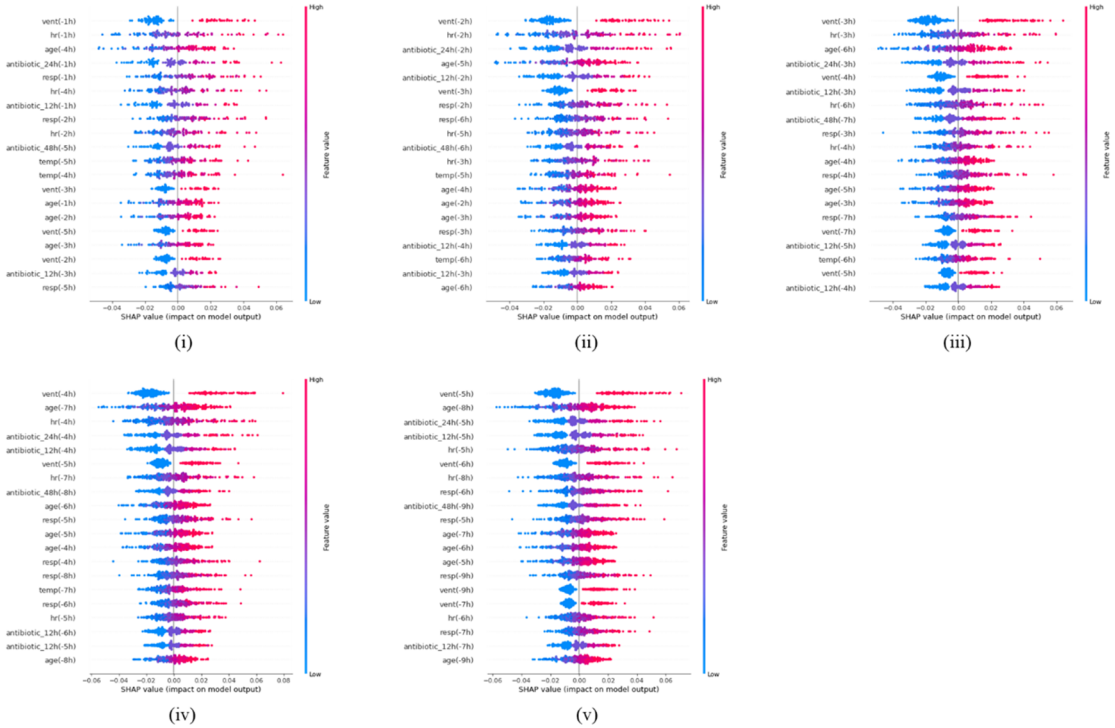
dots that fall on this SHAP value. The color of a dot represents the value of the feature for one sample, with blue indicating low and red indicating high. The names of features on the Y-axis follow the format 'feature (hours before onset)'.

For the models in SEPRES, the feature importance for LightGBM and MLP models was calculated separately in Supplementary Figure 2 and Supplementary Figure 3. The feature importance of LightGBM and MLP models was not consistent overall, although they both yielded high AUCs. For LightGBM models, antibiotics, respiratory rate, total PEEP level, fibrinogen, temperature, net balance, and age were found to be important in most models. For MLP models, age, respiratory rate, ventilation, heart rate, antibiotics, and temperature were found to be important in most models.

Some of these features (antibiotics, respiratory rate, temperature, ventilation, heart rate) were related to the definition of Sepsis-3 or SIRS, while there was also literature arguing for an association between some of these features (respiratory rate,[5] fibrinogen,[6] net balance,[7] and age[8]) and the severity or the mortality of sepsis.



Supplementary Figure 2 Feature importance of LightGBM models in the sepsis early warning module. (i) - (v) are the beeswarm plots of the models detecting sepsis in 1 - 5 hours preceding, respectively.



Supplementary Figure 3 Feature importance of MLP models in the sepsis early warning module. (i) - (v) are the beeswarm plots of the models detecting sepsis in 1 - 5 hours preceding, respectively.

Appendix X Case studies

In the actual operation at Ruijin Hospital, the threshold was increased from the default 0·50 to 0·70 in order to reduce the false alarm rate of sepsis warning. The adjusted specificity was increased to about 0·88, although the sensitivity was reduced to about 0·72, which is shown in Supplementary Table 6.

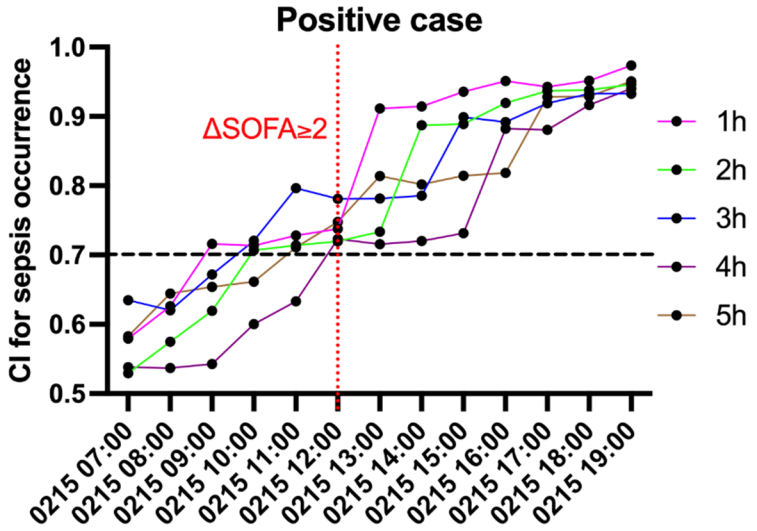
Supplementary Table 6 THE RESULTS OF REAL-TIME DATA WITH THRESHOLD AT 0·70

Preceding hours	Accuracy	AUC	Sensitivity	Specificity
1	0·7480	0·8636	0·6930	0·8861
2	0·7770	0·8789	0·7351	0·8855
3	0·7716	0·8992	0·7265	0·8912
4	0·7747	0·8952	0·7340	0·8852
5	0·7855	0·8858	0·7475	0·8909

Positive case:

On February 14, 2021, a 49-year-old man was transferred to our ICU due to intestinal obstruction and abdominal infection after the pancreaticoduodenectomy (PD) one month ago. After the drainage surgery, the condition of the patient aggravated in the early morning of February 15, 2021, with multiple organ dysfunction. At 12:00 AM, the SOFA score was 13 (ΔSOFA $\geqslant$ 2 within 72 hours). In combination with the suspected infection, the patient was diagnosed as sepsis according to the definition of Sepsis-3. Our sepsis early prediction model has already predicted the incidence of sepsis three hours in advance. At 9:00 AM, the sepsis early prediction model showed that the confidence index (CI) of sepsis incidence in one hour was 0·7161 which was over the prediction threshold 0·70 (Supplementary Figure 4), and the

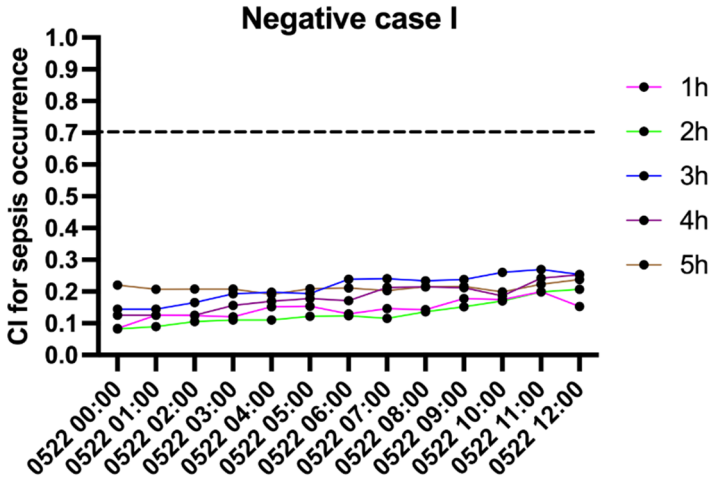
prediction software at the medical terminal UI interface raised the pre-warning prompt. The high CI at 10:00 AM and 11:00 AM indicated the aggravation of the patient. The early prediction of sepsis occurrence by our model effectively guided medical practitioners to appropriately pay more attention to this patient, thereby leading to the early diagnosis of sepsis.



Supplementary Figure 4 The confidence index (CI) of sepsis prediction of the patient at each time node.

Negative case I:

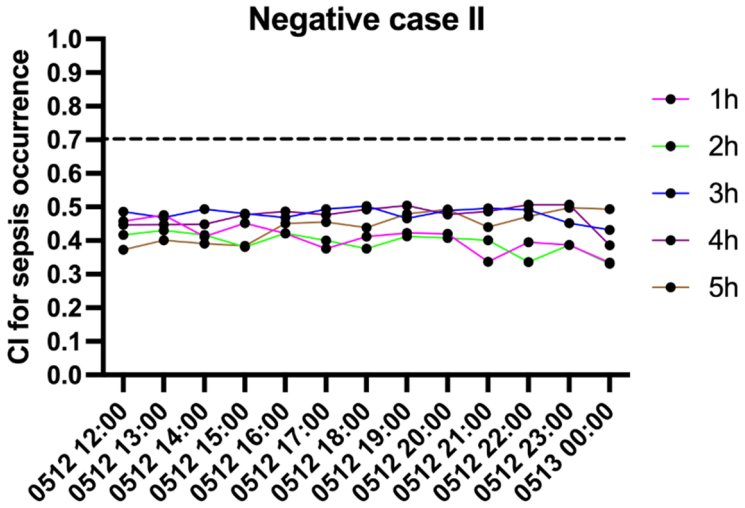
On May 20, 2021, a 45-year-old man was admitted to our hospital due to chronic renal failure (uraemia period with acute exacerbation) in combination with metabolic acidosis and renal failure after renal transplantation as well as severe hypertension. The patient was given RRT hypertension control treatment. During the monitoring in the ICU, the SOFA score was high (7·0), and the condition of the patient was severe due to several comorbidities and complications in combination with uraemia. However, there was no evidence of ΔSOFA ⩾ 2 within 72 hours. Consistently, as shown in Supplementary Figure 5, the CI of sepsis incidence were all lower than the warning threshold 0·70, suggesting no sepsis occurrence after admission. Therefore, this is a negative case.



Supplementary Figure 5 The confidence index (CI) of sepsis prediction of the patient at each time node.

Negative case II:

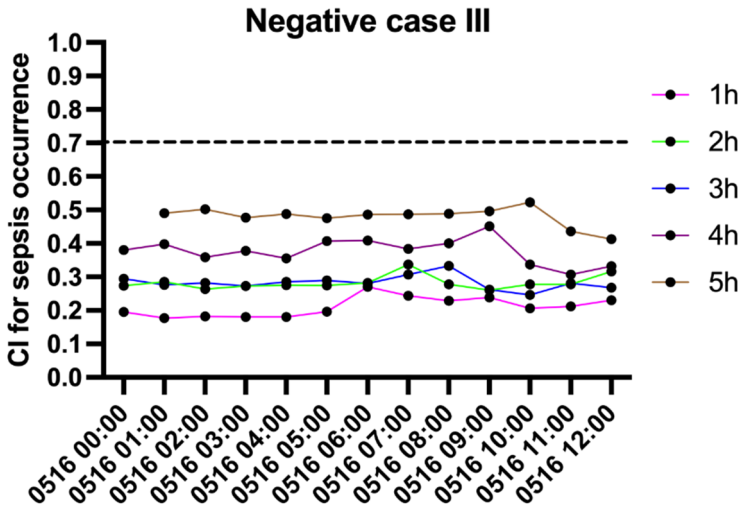
On May 11, 2021, a 58-year-old man was admitted into the ICU due to pancytopenia and bloodstream infection combined with respiratory failure, chronic kidney disease, and severe hypertension and diabetes. After the MDT discussion, the patient was considered to be infection-induced myelosuppression and was given anti-infectious treatment. The patient was diagnosed as sepsis on admission and after the treatment in the ICU, the septic condition was improved, and the vital signs of the patient turned to be stable without the new evidence of ΔSOFA ⩾ 2 within 72 hours. Therefore, the medical terminal interface showed that the CI of sepsis was lower than the warning threshold 0·70 (Supplementary Figure 6), suggesting no sepsis occurrence after admission. Therefore, this is a negative case.



Supplementary Figure 6 The confidence index (CI) of sepsis prediction of the patient at each time node.

Negative case III:

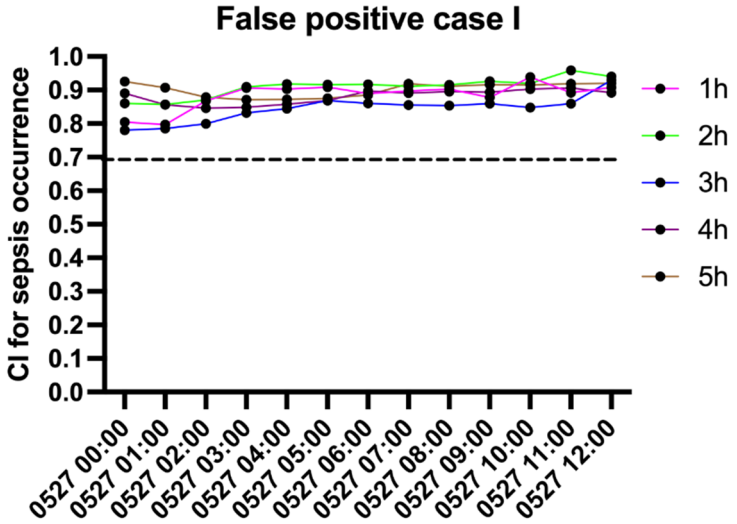
On May 20, 2021, a 69-year-old man was admitted into the ICU due to cardiac insufficiency and pneumonia. After the anti-infectious therapy and organ function maintenance, the condition of the patient was improved, and the vital signs were stable. The SOFA score was stable at 5·0 without the evidence of ΔSOFA ⩾ 2 within 72 hours, suggesting no sepsis occurrence. Consistently, the CI of sepsis predicted by the model was lower than the threshold 0·70, as shown on the medical terminal interface (Supplementary Figure 7), indicating no sign for the warning of sepsis occurrence after admission. Therefore, this is a negative case.



Supplementary Figure 7 The confidence index (CI) of sepsis prediction of the patient at each time node.

False positive case I:

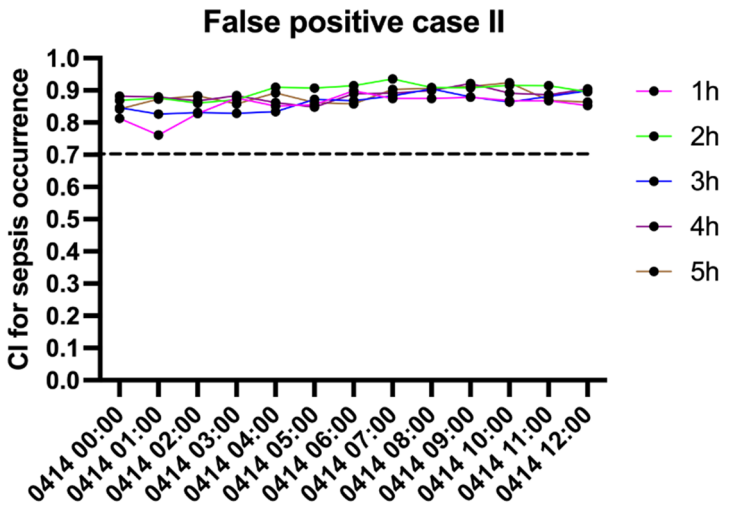
On May 1, 2021, an 84-year-old man was admitted into our hospital due to intracranial space occupying lesion (frontotemporal malignant tumor). The patient was transferred to the ICU due to hospital-acquired pneumonia (klebsiella pneumoniae) and respiratory failure after surgery. The condition of the patient was severe with fluctuations in body temperature during the monitoring period. The SOFA score was stable at 6·0 without the evidence of ΔSOFA ⩾ 2 within 72 hours, suggesting no sepsis occurrence. However, the CI of sepsis incidence in 5 hours predicted by our sepsis early prediction model was over the warning threshold 0·70 (Supplementary Figure 8). Therefore, this is a false positive case.



Supplementary Figure 8 The confidence index (CI) of sepsis prediction of the patient at each time node.

False positive case II:

On January 6, 2021, a 26-year-old man was transferred into the ICU due to cardiogenic shock, hyperthyroid heart disease, and pneumonia. The patient was diagnosed as sepsis on admission and was given ventilation and anti-infectious therapy in the ICU. During the monitoring period, the condition of the patient was very severe, and the body temperature significantly fluctuated. The medical terminal interface showed that the CI of sepsis incidence was over the warning threshold 0·70 (Supplementary Figure 9). However, the SOFA score was stable at 7·0 without the evidence of ΔSOFA ⩾ 2 within 72 hours, and the condition of the patient improved after the control of infection. Therefore, this is a false positive case.

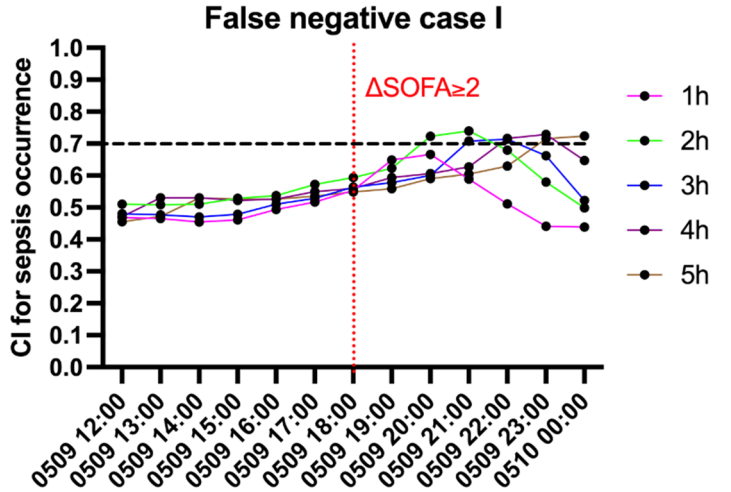


Supplementary Figure 9 The confidence index (CI) of sepsis prediction of the patient at each time node.

The negative control group used during training the sepsis early prediction model was non-septic patients. Therefore, the prediction model may falsely determine the patient that is severely ill to be sepsis. This is a limitation of our present model. To solve this problem, we will analyze the control group by delaminating patients with different severities and further optimize the prediction model.

False negative case I:

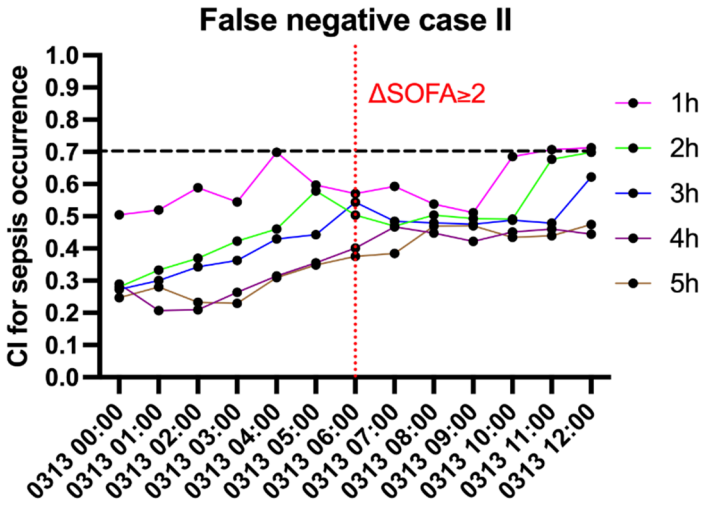
On May 1, 2021, a 63-year-old man was admitted into our hospital due to sellar tumor. On May 8, 2021, during the transnasal transsphenoidal resection of pituitary adenoma, the patient had hemorrhagic shock and hypoxic-ischemic encephalopathy. The patient was transferred into the ICU and was given ventilation and anti-infectious therapy. The SOFA score showed an increase from 6·0 to 9·0 (ΔSOFA ⩾ 2 within 72 hours) at 06:00 PM on May 9, 2021. In combination with the evidence of infection, the patient was diagnosed as sepsis. However, as shown on the medical terminal interface (Supplementary Figure 10), the CI of sepsis incidence was below the warning threshold 0·70. Therefore, this is a false negative case.



Supplementary Figure 10 The confidence index (CI) of sepsis prediction of the patient at each time node.

False negative case II:

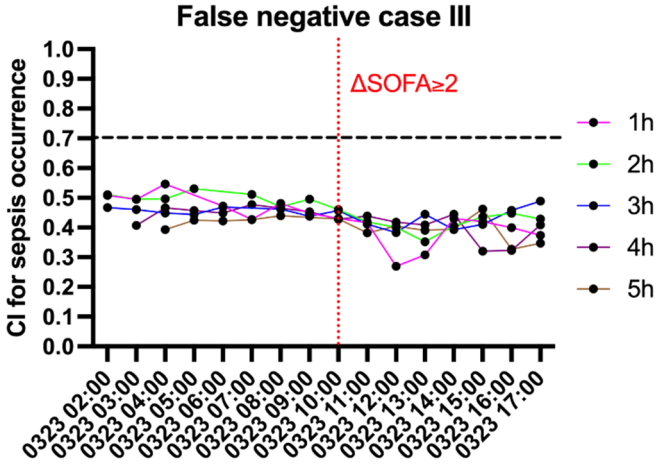
On March 11, 2021, a 46-year-old man was admitted into the ICU due to severe acute pancreatitis and abdominal infection. During the monitoring period, the patient had an obvious fluctuation in body temperature, and at 06:00 AM on March 13, 2021, there was an obvious increase in SOFA score (ΔSOFA ⩾ 2 within 72 hours), indicating the occurrence of sepsis. However, the CI of sepsis incidence was below the warning threshold 0·70 (Supplementary Figure 11). Therefore, this is a false negative case.



Supplementary Figure 11 The confidence index (CI) of sepsis prediction of the patient at each time node.

False negative case III:

On March 15, 2021, a 55-year-old man was admitted into our hospital due to severe acute pancreatitis. After debridement and drainage of necrotizing pancreatitis on March 22, 2021, the patient was transferred into the ICU and was given anti-infectious therapy. During the monitoring period, the patient had an obvious fluctuation in body temperature and at 10:00 AM on March 23, 2021, there was an obvious increase in SOFA score (ΔSOFA ⩾ 2 within 72 hours), indicating the occurrence of sepsis. However, the CI of sepsis incidence was below the warning threshold 0·70 (Supplementary Figure 12). Therefore, this is a false negative case.



Supplementary Figure 12 The confidence index (CI) of sepsis prediction of the patient at each time

node.

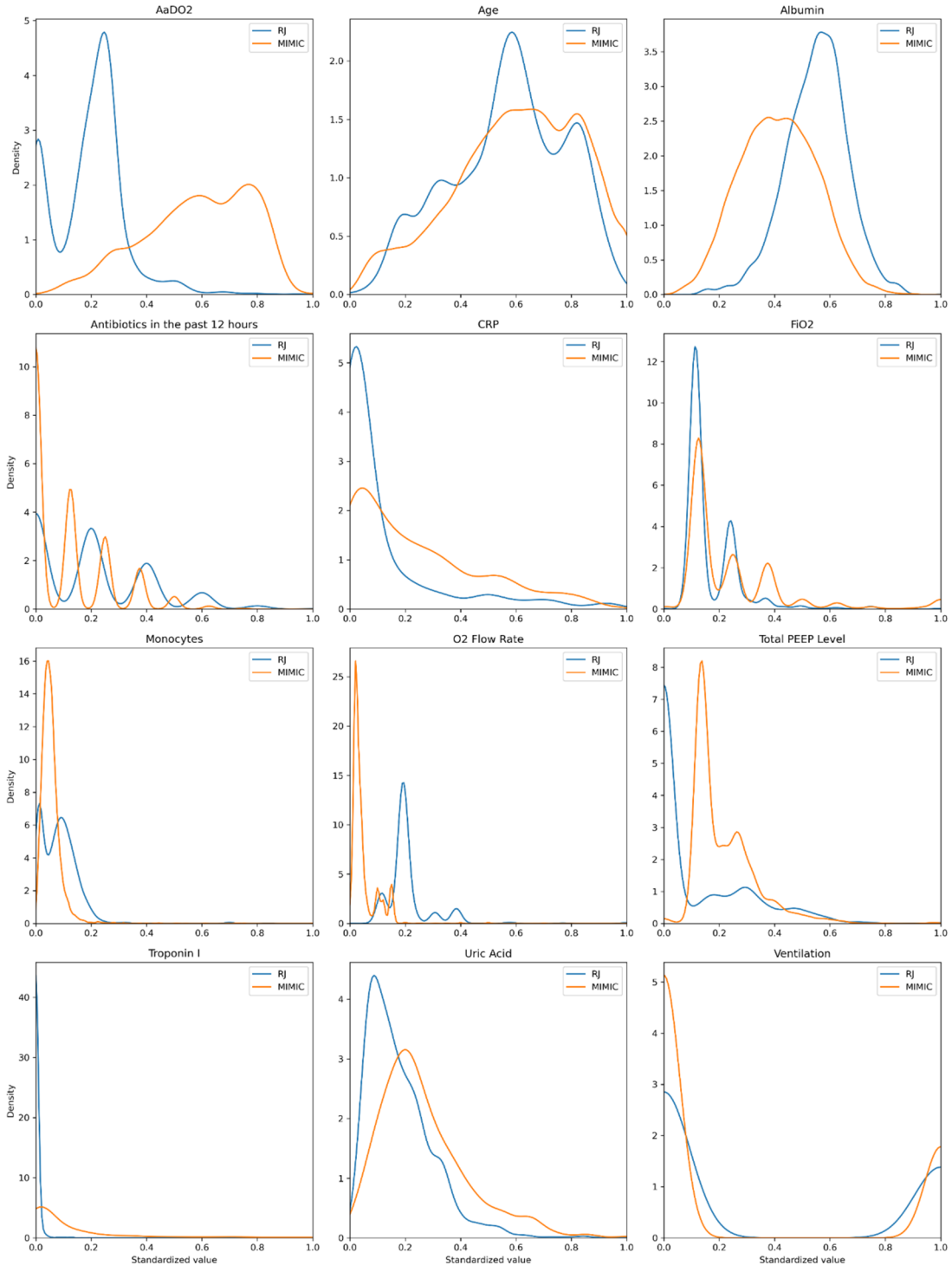
We concluded that the ICU duration for all these three false negative cases was relatively short, therefore the data collected for prediction was limited, which may lead to inaccurate prediction for sepsis occurrence. As supporting evidence, there were only 17 positive cases with an onset within 2 days in the RJ historical dataset, which may not be sufficient for the model to learn positive cases in the absence of enough data.

As a further experiment, we selected cases in the RJ database that were screened out due to too few valid variables, and the performance on these cases is shown in Supplementary Table 7. It can be seen that the sensitivity produced a substantial decrease in sensitivity in contrast to a slight decrease in specificity. This suggests that when there is insufficient valid data for a case, the model will tend to produce lower CI of sepsis incidence, resulting in false negative cases. This may need to be addressed by the inclusion of more positive cases with fewer valid variables, as well as data augmentation.

Supplementary Table 7 THE RESULTS ON CASES WITH FEW VALID VARIABLES

Preceding hours	Accuracy	AUC	Sensitivity	Specificity
1	0·6325	0·6492	0·3150	0·9500
2	0·6283	0·6740	0·3167	0·9400
3	0·6338	0·6680	0·3300	0·9375
4	0·6350	0·6706	0·3180	0·9520
5	0·6283	0·6865	0·3050	0·9517

Appendix XI Differences in features between MIMIC dataset and RJ dataset



Supplementary Figure 13 Feature distributions with relatively large differences between the MIMIC dataset and the RJ dataset. Each subplot shows the density functions of a feature on two datasets obtained by kernel density estimation using the Gaussian kernel function.

Reference

1 Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An interpretable machine

learning model for accurate prediction of sepsis in the ICU. *Crit Care Med* 2018; **46**: 547.

2 Moor M, Horn M, Rieck B, Roqueiro D, Borgwardt K. Early recognition of sepsis with Gaussian process temporal convolutional networks and dynamic time warping. Machine Learning for Healthcare Conference; 2019: PMLR.

3 Hong LK, Wogan G, Vacca L, Tidor B, inventors. Peach IntelliHealth Pte Ltd., assignee. System and method for predicting sequential organ failure assessment (sofa) scores using artificial intelligence and machine learning. United States Patent 20190259499; 2019.

4 Silva Á, Cortez P, Santos MF, Gomes L, Neves J. Rating organ failure via adverse events using data mining in the intensive care unit. *Artif Intell Med* 2008; **43**: 179-93.

5 Kenzaka T, Okayama M, Kuroki S, et al. Importance of vital signs to the early diagnosis and severity of sepsis: association between vital signs and sequential organ failure assessment score in patients with sepsis. *Intern Med* 2012; **51**: 871-6.

6 Matsubara T, Yamakawa K, Umemura Y, et al. Significance of plasma fibrinogen level and antithrombin activity in sepsis: a multicenter cohort study using a cubic spline model. *Thromb Res* 2019; **181**: 17-23.

7 Brotfain E, Koyfman L, Toledano R, et al. Positive fluid balance as a major predictor of clinical outcome of patients with sepsis/septic shock after ICU discharge. *Am J Emerg Med* 2016; **34**: 2122-6.

8 Martin GS, Mannino DM, Moss M. The effect of age on the development and outcome of adult sepsis. *Crit Care Med* 2006; **34**: 15-21.